# Nonparametric Statistical Modeling of Recurrent Events: A Bayesian Approach

Andriy Andreev

Division of Biometry
Rolf Nevanlinna Institute

Faculty of Science
University of Helsinki

Academic Dissertation for the Degree of Doctor of Philosophy

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium III, Porthania, on June 30th, 2000, at 12 a.m. noon

# Nonparametric Statistical Modeling of Recurrent Events: A Bayesian Approach

Andriy Andreev

Division of Biometry
Rolf Nevanlinna Institute
University of Helsinki

Research Reports A32
June 2000

Rolf Nevanlinna Institute
Res Inst Math Stat & Comp Sci
P.O. Box 4 (Yliopistokatu 5)
FIN-00014 University of Helsinki, Finland

ISBN 952-9528-59-0
ISBN 952-91-2274-8 (PDF version)

YLIOPISTOPAINO
HELSINKI 2000

# Acknowledgements

I wish to express my deepest gratitude to Professor Elja Arjas for his guidance and support on my way to professionalism. He encouraged me to become an independent researcher all the way through the years of working on this Ph.D. thesis.

My thanks are due to Dario Gasbarra, with whom I shared office and ideas for many years. I am grateful to Dr. Mervi Eerola for intense collaboration which resulted in submitting article to the leading statistical journal.

I would also like to express my gratitude to all people at the Department of Mathematics, University of Oulu, and at Rolf Nevanlinna Institute, University of Helsinki, for establishing working environment. Personally, I want to mention Pekka Kangas and Matti Taskinen who familiarized me with computer software.

My thanks are also to Prof. Ø. Borgan and to Prof. P. Volf for writing their review reports on the manuscript of this thesis. They enriched my view on the work. I wish also to thank Professor Hannu Oja who kindly has accepted to be my opponent at public defense.

Very special thanks I address to my parents and friends for their love, beauty and patience: values which are so important for creative human relations. I also wish to thank my teachers at Kiev University for building my research potential.

There are many other people, whose help was indispensable at different stages of research. For their contribution I refer to the Acknowledgment sections of the original publications.

Financial support from Academy of Finland, Rolf Nevanlinna Institute, CIMO, ComBi graduate school, and Rector of University of Helsinki are gratefully acknowledged. Permissions to reprint original articles have been granted by LIDA and Oxford University Press.

Helsinki, May 2000
Andriy Andreev

"Mathematicians always want their mathematics to be pure, that is, strict and provable, wherever possible. However, the most interesting problems can not usually be solved in this manner. Therefore, it is very important that a mathematician should be able to find approximative (not necessary strict but effective) ways of solving such problems."
The last interview of A.N. Kolmogorov with a documentary film maker, A.N. Marutyan.

# Contents

# 1 Introduction

This thesis focuses on applications of nonparametric Bayes theory to correlated multivariate failure time data. A Counting Processes framework is used for model building. The entertained models are built in order to fit real data sets from medical applications. The solution of inferential and goodness-of-fit problems is made feasible by application of Markov chain Monte Carlo (MCMC) methods.

A nonparametrically defined piecewise constant intensity model for acute ear infections (AOM), involving both fixed and time dependent covariates, is introduced in the second paper. In the third paper, the notion of predictive distribution is used, firstly, in an attempt to draw causal inferences, and secondly, in an assessment of the performance of the applied nonparametric Bayesian model. The results of formal tests developed on the basis of estimates of predictive distributions led to changes in the model constructed in the second paper.

The fourth paper attempts to build a new model for the same underlying process of AOM based on a new, more elaborate data set. Two new concepts of "antibody level" and of "being a carrier of infection" (a binary process indicating exposure to the occurrence of AOM) are introduced in order to model the intensity of AOM. Understanding of interactions between these three processes is a practically motivated problem. It corresponds to "looking inside" the real processes of being exposed to infection, of being a carrier of infection and then investigating the protective reaction in terms of levels of specific antibodies, and finally, the exposure need to acquire AOM conditionally at the level of antibodies and some environmental factor. Conceptually, this paper is a direct continuation of the previous two articles.

Problems of Bayesian density estimation are discussed in the first paper. Though the reported results may not find any apparent use in applications, they could serve as an introduction to the theoretical aspects of applied modeling. The main result of the paper states weak sufficient conditions which the likelihood function should satisfy in order to obtain a tractable posterior density approximation based on algorithmically constructed piecewise constant approximation of the prior.

The following three sections contain preliminaries of general methodological concepts common to the original papers constituting this thesis. A short introductory overview on statistical model building/checking for survival data is given in the other sections. This is the area where the models of this thesis find their methodological and practical applications. A short general discussion concludes the introductory part.

# 2 Bayesian approach to modeling data

Bayes' theorem is one of the widely accepted bases for statistical inference. Opinions as to the value of Bayesian approach have jumped from acceptance to rejection and vice versa since the introduction of the original formula in 1763. During periods when

alternative methods gave solutions, Bayesian results were viewed with suspicion. When new problems were stated or classical approaches came to their limits, there was always renewed respect for Bayesian methods.

The technical result we now know as *Bayes theorem*, in its simplest form, could be stated as follows

$$P(H|data) = \frac{P(data|H)P(H)}{P(data)}, \tag{1}$$

with $P(H)$ standing for the prior belief about hypothesis $H$ before obtaining *data*. Hypotheses always comprise subjective beliefs about values of parameters of interest.

Like any equality, formula (1) provides a statement that the left-hand side of the equality must equal the right-hand side. Applied interest usually lies in the interpretation. At the heart of the controversy is the issue of the interpretation of probability: objective or subjective, and the legitimacy of basing a scientific theory on the latter (see Bernardo and Smith (1994) for a detailed discussion of the issue). This is done by assigning prior distribution to the parameters of interest. A thorough account and defense of this point of view are given in de Finetti's (1974,1975) two-volume *Theory of probability*.

In the above formulation, there are two objects which should be specified in an attempt to describe any process using Bayes modeling: prior and likelihood. The Bayes modeling adds a prior specification $P(H)$ to the likelihood. Inference is based on the posterior distribution $P(H|data) \propto L(H, data)P(H)$.

The choice of an appropriate prior is always a very delicate procedure in Bayesian analysis. Often people choose to obey the simple rule of selecting a "non-informative prior", (see e.g. Lindley (1965)). The idea of a non-informative prior distribution, representing "ignorance" and "letting the data speak for themselves", is often regarded as synonymous with providing objective inferences. This search for "objectivity" is rather misleading, since there is no known form for a "true" non-informative prior. One can rather speak about "minimally-informative" prior specifications (see Bernardo and Smith (1994), Chapter 5.6.2 for a more detailed discussion on issues of prior ignorance).

Another specific feature of Bayesian approach could be stated as follows. While other inferential theories rely heavily on strict model assumptions like normality, Bayesians could allow greater emphasis on scientific interest and less on mathematical convenience.

The importance of this advantage becomes obvious when one compares the price of data gathering to that of data analysis. Data gathering may last for many years and is often conducted by institutions different from statistical research centers. This process is costly and requires the establishment of the same standards of recording data for all sources. Often recorded data sets are incomplete or do not contain enough information to perform statistically acceptable inferential procedures. Added to this, it seems to be always quite problematic to obtain data sets for independent statistical research.

That is why data analysis partially plays a complementary role to that of data gathering. Often models could indicate/predict certain characteristics which are important for the adequate description of the process of interest but which were for some reason neglected in data gathering. These considerations often lead to changes in data recording.

The ultimate goal of statistical modeling in this respect is to develop generic techniques/recipes for collecting and then analyzing data sets of a certain type. In order to do this, one nowadays has to rely on computer power and algorithmic procedures. MCMC techniques, described in the fourth chapter, make a Bayesian perspective for developing such procedures more promising.

# 3 Point Processes framework

Bayesian methodology is not the only common feature of the collection of papers presented in the thesis. A second key notion is the theory of Point Processes. All three applied models are built within this unifying framework of conducting event history analysis. The latter means the study of a collection of individuals, each moving among a finite number of states. In order to exemplify the notion, one can think about the simplest possible situation of moving from the state "alive" to the state "dead". This example forms the basis of survival analysis. Models presented in this thesis deal with transitions such as healthy/sick.

The Counting Process associated with the Point Process is characterized by a dynamic process (intensity), and the special pattern of incompleteness of observations (right-censoring or left-truncation in our case). This characterization is an application of the well known Doob-Meyer decomposition theorem. Having defined the intensity process, one is interested in estimation of its parameters. Inferential procedures in this framework first were presented in Aalen (1975), and turned out to be very fruitful. For further developments, see Andersen *et al.* (1993).

We state here the basic features of the theory employed in the three applied papers of this thesis.

**Definition 1.** A counting process is a stochastic process $\{N(t) : t \geq 0\}$ adapted to a self-exciting filtration $\{\mathcal{F}_t : t \geq 0\}$ with $N(0) = 0$ and $N(t) < \infty$ a.s., and whose paths are with probability one right-continuous, piecewise constant, and has only jump discontinuities, with jumps of size one.

Having made the choice of the way of modeling, let us turn back to inferential aspects of the study. In what follows, we define a likelihood function for the class of models based on counting processes, and discuss its properties.

For the selection of the likelihood function we followed a (see e.g. Andersen *et al.* (1993), Chapter 2.7 or Fleming and Harrington (1991)) well developed theory leading to a Poisson type of likelihood. The argumentation is based on Jacod's Formula for the Likelihood Ratio. We state the appropriate result here in the case

of the existence of absolutely continuous compensators.

**Theorem.** (simplified Jacod's) Let $\mathcal{F}_t = \mathcal{F}_0 \bigvee \sigma\{N(s) : s \leq t\}$. Let $P$ and $\tilde{P}$ be two probability measures on the filtered probability space under which $N$ has $P$-a.s. absolutely continuous compensators $\Lambda$ and $\tilde{\Lambda}$, respectively. Suppose $\tilde{P}$ is absolutely continuous w.r.t. $P$, written $\tilde{P} << P$. Then

$$\frac{d\tilde{P}}{dP}|_{\mathcal{F}_t} = \frac{d\tilde{P}}{dP}|_{\mathcal{F}_0} \frac{\prod_s [\tilde{\lambda}(s)^{\triangle N(s)}] exp\{-\tilde{\Lambda}(t)\}}{\prod_s [\lambda(s)^{\triangle N(s)}] exp\{-\Lambda(t)\}}. \tag{2}$$

One can easily see that products have as many terms as the underlying counting process $N(t)$ has jumps. In the applications one can consider likelihood ratios of form (2) formed by taking Radon Nikodym derivatives w.r.t. one fixed reference measure. Likelihoods are only needed up to a proportionality factor. The numerator in formula (2) has a Poisson form and it is used as the likelihood in our applications. The denominator is written w.r.t. a fixed reference measure and treated as a constant.

# 4 Markov chain Monte Carlo

Another common general concept, already mentioned earlier, is MCMC. This is the technique which made computational issues of the presented thesis feasible. It forms a basis for calculation of integrals w.r.t posterior densities $p(\theta|data)$, where $\theta$ is a vector of parameters. Expression $p(\theta|data) = \frac{p(\theta)p(data|\theta)}{\int p(\theta)p(data|\theta)d\theta}$ defines a posterior density of $\theta$. Most of quantities legitimate for Bayesian inference: moments, probabilities, etc., can be expressed as follows:

$$E[f(\theta)|data] = \frac{\int_\Theta f(\theta)p(data|\theta)p(\theta)d\theta}{\int_\Theta p(data|\theta)p(\theta)d\theta}.$$

The integrations in this expression have until recently been the source of computational difficulties, in particular if the parameter space $\Theta$ has a high dimension. MCMC is an available computational technique which often offers a unified framework for solving this problem.

Though the last sentence sounds very optimistic, and MCMC techniques really allow to perform the calculations for large models, one should not forget that the main problem always was and will be in the modeling part. What matters after all is the skill of the model builder. MCMC gives him/her a tool to realize more complex beliefs about a "true model" in practice.

The first applications of MCMC were in the statistical physics literature with further developments in spatial statistics and image analysis. Applications in the context of Bayesian inference are much more recent (see e.g. W.R. Gilks *et al.* (1996)).

The form of the posterior density $p(\theta|data)$ often does not allow for direct independent sampling. MCMC builds a Markov chain with this posterior as the invariant

distribution instead. Having sampled from this chain, one can estimate the characteristics of the posterior. How well this strategy works in a particular problem depends on certain characteristics of the Markov chain. Under some regularity conditions, the following approximation is valid

$$\int_{\Theta} f(\theta) d\pi(\theta) \approx \frac{1}{n} \Sigma_{i=1}^{n} f(\theta_i), \tag{3}$$

where $\{\theta_i, i = 1, ..., n\}$ is a dependent sample forming a Markov chain with ergodic distribution $\pi(\theta) = p(\theta|data)$.

The question is then how to construct a Markov chain such that its stationary distribution is precisely the posterior distribution. It happens to be an easy task due to the Metropolis-Hastings algorithm (see e.g. Tierney (1994)). Suppose that $\pi(\cdot)$ is a distribution we want to sample from (our posterior). At each time $t$, the next state $X_{t+1}$ of the chain is chosen by first sampling a candidate $Y$ from a proposal distribution $q(\cdot|X_t)$. The candidate point $Y$ is then accepted with probability $\alpha(X, Y)$ where $\alpha(X, Y) = min(1, \frac{\pi(Y)q(X|Y)}{\pi(X)q(Y|X)})$. If the candidate point is accepted, the next state becomes $X_{t+1} = Y$. Otherwise, the chain remains in the same state for one more unit of time $X_{t+1} = X_t$.

The sampling algorithms adopted in this thesis are all based on the above depicted basic idea. "Birth and death" version of the Metropolis-Hastings algorithm for a point process which admits a density w.r.t. the Poisson measure (see Geyer and Møller (1994)) is used in the fourth paper.

# 5 On one classification of survival data

There are two sampling designs which are commonly used in longitudinal medical studies. The first design uses the initial occurrence of an event as the enrollment criterion, and repeated occurrences of the same event are observed within a considered time period. In this case the occurrence time of the initial event is defined as the time origin. In the second design, subjects are sampled from a target population and recurrent events are observed during a follow-up period. Thus it is possible that no events occurred during the follow-up period. In this study we have develop statistical methods focusing on the second design.

Generally, applied statistical models in longitudinal medical studies are built in order to describe relations between the occurrence of a disease and explanatory variables (covariates such as treatments, design variables, etc.). We follow here one of many possible classifications of survival data encountered by statisticians in medical studies (see Sinha and Dey (1997)): univariate survival data, multiple event-time data, multivariate survival data.

Univariate survival data occur in case when every subject of the study can experience the event at most once, such as death of an individual or failure of a machine. We speak about multiple event-time data when subject experiences an event more than

once. When some survival times are related, such as in litter-matched studies, one speaks about modeling of multivariate survival data.

Let us outline one basic approach to modeling data according to the classification introduced above. One of the most widely used methods for analyzing *univariate survival data* utilizes Cox's (1972a) semiparametric proportional hazards model $\lambda(t|\theta) = \lambda_0(t)exp\{\beta^T\theta\}$, where $\lambda_0(t)$ is an unknown, nonparametrically modelled baseline hazard function, $\beta$ is an unknown vector of regression coefficients, and $\theta$ is a covariate vector (the vector of explanatory and design variables).

A semiparametric model of *multiple event-time data*, the so-called proportional intensity model, was presented for example in the work of Sinha (1993). It employs a conditional Poisson process with the conditional intensity function

$$\lambda_i(t|\theta, Z_i) = Z_i\lambda_0(t)exp\{\beta^T\theta\}, \tag{4}$$

given the unobserved "frailty" $Z_i$ and the observed covariate $\theta$. The frailty random variables $Z_i's$ account for the individually specific dependencies among interevent times of recurrent ear infections, and for factors, not captured by other model parameters (see the appropriate parts of the second and fourth papers).

These models are useful when the focus is on the influence of covariates (or treatments) and on the heterogeneity among subjects related to the rate of occurrence of events in each subject (see Vaupel, Manton, and Stallard (1979)). See also Oakes (1992) for several other models proposed for multiple event-time data in the context of frequentist inference.

*Multiple event-time data* group is the area where models of this thesis find their applications. We build models of the type, specified in formula (4). The differences come in choosing other than exponential link functions or modeling more than one part of the model nonparametrically. For example, in the fourth paper we model $\alpha_t$ nonparametrically. It stands for the time-varying regression coefficient of antibody level, and it is modelled in the same way as the baseline function. Moreover, observed covariates, denoted by $\theta$ in formula (4), are basically partially observable and come into the model using the idea of data augmentation (see Tanner and Wang (1987)).

Semiparametric models for multivariate survival data can be classified into two groups: conditional and marginal. Conditional models are used mainly in semiparametric Bayesian and likelihood methods of analyses. They induce a correlation structure between related events (in our case infections) through an unobserved random process. Given this random process, the infections within a patient are independent. Marginal models are more popular in generalized estimating equations (GEE) methodologies.

# 6 Model checking aspects

One common problem present in modeling of multiple event-time data is the complexity of suitable methods to verify the modeling assumptions. In practice, one usually needs to address two problems, that is, model adequacy and model selection. The latter is the problem of selecting the "right form" of the model, and the former answers the question whether the chosen model works properly.

Model adequacy is of particular interest for us in this thesis. The third paper tackles this problem by offering both a graphical-check technique and a formal quantitative test for the evaluation of model fit. In what follows we give a short overview of the alternatives for checking model adequacy discussed recently in the literature.

The literature for model adequacy in a Bayesian framework does not seem to be rich. The formal Bayesian model adequacy criterion (as in Box 1980) proposes that the marginal prior predictive density is to be evaluated at the times of observations. Large values of density support the model; small values do not.

The posterior predictive model checking (motivated by Gelman, Meng, and Stern (1996)) goes as follows. Let $y_{obs}$ be the observed data, $\theta$ be the vector of unknown parameters in the model. We assume that have already obtained draws $\theta_1, ..., \theta_n$ from the posterior distribution, possibly using Markov chain simulations. We now simulate $N$ hypothetical replications of the data $y_1^{new}, ..., y_N^{new}$, where $y_i^{new}$ is drawn from the sampling distribution of $y_{obs}$ given the simulated parameters. Thus $y^{new}$ has distribution $P(y^{new}|y_{obs}) = \int P(y^{new}|\theta)P(\theta|y_{obs})d\theta$. Generating a sample $y^{new}$ adapts the old idea of comparing data to simulations from a model, with the Bayesian twist that the parameters of the model are themselves drawn from the posterior distribution.

If the model is reasonably accurate, the simulated data should look similar to the observed data $y_{obs}$. Formally, one can compare the data to the predictive distribution by first choosing a discrepancy variable *test statistic* $S(y, \theta)$ which will have an extreme value if the data $y_{obs}$ are in conflict with the model. Then a $p$-value can be estimated by calculating the proportion of cases in which the simulated discrepancy variable exceeds the realized one: estimated $p$-value $= \frac{1}{N}\Sigma_{i=1}^{N} I(S(y_i^{new}, \theta_i) \geq S(y_{obs}, \theta_i))$, where $I(\cdot)$ is the indicator function.

In practice, we can often visually examine the posterior predictive distribution of the test statistic and compare it to the realized value. If the test statistic depends only on data and not on the parameters $\theta$, we can plot a histogram of posterior predictive simulations $S(y^{new})$ and compare it to the observed value $S(y_{obs})$.

The discrepancy variable can be any function of data and parameters. It is the most useful to choose a test statistic which measures some aspect of the data that might not be accurately fitted by the model. A model does not fit the data if the realized values for some meaningful test statistics cannot reasonably be explained by chance. This is the case when the tail-area probability is close to 0 or 1. The $p$-values, defined as $P(S(y^{new}, \theta) \geq S(y_{obs}, \theta)|y_{obs})$ are actual posterior probabilities and can therefore

be interpreted directly. A lack of rejection should not be interpreted as "acceptance" of the model, but rather as a sign that the model adequately fits the aspect of the data being investigated.

Major failures of the model can be addressed by expanding (we did this by introducing the mixture model for the frailty parameter) the model. Lesser failures may also suggest model improvements or might be ignored in the short term if the failure appears not to affect the main inferences.

Unfortunately, such checking procedures seem to be technically very difficult, or even unfeasible, for most of complex survival data problems. This is a formidable task to find appropriate test statistics and then improve the model so that it gives reasonable results for all of them. Another reason for using other criteria comes in applications with censored survival data. One can not directly apply the above mentioned rules to the censored observations (both, prior and posterior predictive density tests).

Bayesian exploratory analysis diagnostics methods for model adequacy is another option considered in the survival data literature. It is mainly based on appropriate residuals (see Andersen *et al.* (1993) for examples). When one has evaluated residuals, one builds a test which summarizes their values. These tests could be graphical (like TTT or Q-Q-plot techniques), or numerical, based on a knowledge of the expected posterior distributions. They can neatly incorporate incomplete (like right censored) data structures. We adapted one version of this type methods in the third paper. The argumentation is based on the martingale central limit theorem for counting process. For details and list of necessary assumptions about the structure of the data, see the appropriate chapters of the paper.

Another graphical approach developed for model adequacy is based on the conditional predictive ordinate (CPO). Formally, the CPO for the $i^{\text{th}}$ observation $y_i$ is defined as the cross-validated density $f(y_i|y_{(i)})$, where $y_{(i)}$ denotes the whole data set except for the $i^{\text{th}}$ observation. Repeating CPO for all observations, one can use results for the model diagnostics. A large CPO indicates agreement between the observation and the model.

# 7 Model Comparison Aspects

Model comparison is another fundamental issue of model determination. Indeed, it comes up naturally when one has built, let us say, two models giving adequate fit to the data. Then the obvious question to ask is "How to choose the better one of the two?". This could be accomplished in two ways: by testing each model separately and then selecting the better one for chosen criteria or by joint/simultaneous testing. Let us note that one can think about much more general and analytically ambitious version of the formulated problem, i.e. of finding an optimal model in some class (possibly infinite dimensional) of parametric/nonparametric models. We restrict ourselves here to the simpler problem of finding the "better one" of two.

This thesis does not contain much material on the comparison of models. In what follows, we rather give a perspective for future research. Up to the present there have been several studies performed on the same/similar data sets (see e.g. Oja *et al.*, etc.). It would be desirable to understand what aspects of data reflected better in each of these studies. Is there any model which performs better than others, making them obsolete? Fortunately for solving this problem, sampling-based methods using MCMC techniques not only enable the desired inferences to be performed, but also facilitate model determination. The first test problem which could be formulated is to compare, using the techniques described below, the mixture-frailty model of the third article with the one from the second paper.

This short overview of techniques for model comparison closely follows the paper of Gelfand and Dey (1994). That paper suggests a unifying framework for introducing/using different versions of Bayes factors (BF) developed in the earlier literature. In doing this, the notion of the generalized predictive density is introduced.

All complex models in a Bayesian context may be viewed as the specification of a joint distribution of unobservables (model parameters, missing data or latent variables) and data. Regardless of the structure, it is this joint density whose performance must be examined w.r.t. model determination. There are several techniques which try to elicit the problem. The scope of activity on model selection/comparison is revealed in a series of recent works on Bayes factor (see e.g. Gelfand and Mallick (1995)), cross-validation (Gelfand and Dey (1994)), intrinsic Bayes factors (Berger and Pericchi (1996)), and posterior Bayes factors (Aitkin (1996)). In all these techniques the choice is made by reducing each model to a single summary number and then comparing these numbers.

In what follows we also give a short outlook of how MCMC can be employed for evaluating asymptotic behaviour of estimators arising from one version of predictive density closely related to the one used in the third paper.

## 7.1 Formulation of the Problem

Let us start by outlining briefly one classical approach to the problem of model choice. In what follows, we make a choice between two parametric models denoted interchangeably by joint density $f(y|\theta_i; M_i)$ or likelihood $L(\theta_i; y, M_i)$, $i = 1, 2$, where *y(data)* is $n \times 1$ and $\theta_i$ is $p_i \times 1$. Following the Neyman-Pearson theory, let $H_1$ (*data* arise from model $M_1$) be the null hypothesis, where $M_1$ is nested in the full model $M_2$, where $M_1$ is nested in the full model $M_2$. The *likelihood ratio test* then rejects $H_1$ if $\lambda_n < c < 1$ where

$$\lambda_n = L(\hat{\theta}_1; y, M_1)/L(\hat{\theta}_2; y, M_2).$$

Assuming that the models are regular, i.e. $p_i$ is finite as $n \to \infty$, under mild conditions, $-2log(\lambda_n)$ is approximately distributed as $\chi^2_{p_2-p_1}$ under $H_1$. The drawback is that the likelihood ratio test is inconsistent. This follows from the following simple

consideration

$$lim_{n\to\infty}P(\text{choose } M_2|M_1 \text{ is true}) = lim_{n\to\infty}P(\lambda_n < c|M_1 \text{ is true})$$
$$= P(\chi^2_{p_2-p_1} > -2log(c)).$$

This probability is positive. If $M_1$ is the reduced model and $M_2$ is the full model, the likelihood ratio test gives too much weight to the full model. To account for this deficiency, there have been developed many versions of so-called *penalized likelihood tests* in the form $L(\hat{\theta}_i; y, M_i) - k(n, p_i)$. Let us now come to Bayes analogue of the likelihood ratio test.

Bayesian inference is based on the posterior distribution $\pi(\theta|y) \propto L(\theta, y)\pi(\theta)$, where $\pi(\theta)$ is the assigned prior. The case when $L$ is held fixed and $\pi$ varied is referred to as Bayesian robustness (see Berger (1985)).

We will be interested to draw parallels with the above described classical situation by varying the likelihood function and keeping the prior fixed. Let $w_i$ be a prior probability of $M_i$, $i = 1, 2$, and $f(y|M_i)$ is the *Predictive Distribution* for model $M_i$, i.e. $f(y|M_i) = \int f(y|\theta_i, M_i)\pi(\theta_i|M_i)d\theta_i$. If $y_{obs}$ denotes the observed data then we use the Bayes factor of $M_1$ w.r.t. $M_2$ in the following form

$$\text{BF} = \frac{w_1 \cdot f(y_{obs}|M_1)}{w_2 \cdot f(y_{obs}|M_2)}. \tag{5}$$

Having calculated (5), why does one have to look for alternatives? The problems are similar to those of the likelihood ratio test.

The BF, in the nested model case, under usual regularity conditions tends to infinity as $n \to \infty$, i.e. regardless of the data, as $n$ grows large, model $M_1$ will be chosen. In other words, the conclusion is rather contradictory to the one from the likelihood ratio test. In order to account for this theoretical problem, several versions of *BF* were proposed. In the next subsection we introduce the notion of general predictive density which allows us to classify different types of *BF*.

## 7.2 A Short Classification of Bayes Factors

The underlying suggestion is to adopt a broader notion of predictive density. According to Gelfand and Dey (1994), "predictive density arises by averaging a density defined over some portion of the sample space (arising from the likelihood) w.r.t. a distribution on the parameter space (arising from a data-based updating of the prior)".

Let us assume that data $y = (y_j,\ j \in J_n)$, $J_n = \{1, ..., n\}$ have density $f(y_j|\theta_i, M_i)$, $i = 1, 2$. In what follows we also assume that $y_k \perp y_l$ given $\theta$, for any $k \neq l$, where $k$ and $l$ could be vectors. Following Gelfand and Dey (1994), let us define

$$L(\theta_i; y_S, M_i) = \prod_{j=1}^{n} f(y_j|\theta_i, M_i)^{d_j},$$

where $d_j = 1$ if $j \in S$ or $d_j = 0$ if $j \notin S$. Then, a predictive density which averages the joint density of $y_{S_1}$ ($S_1 \subseteq J_n$) w.r.t. the prior of $\theta_i$ updated by $y_{S_2}$ ($S_2 \subseteq J_n$) can be written in the following general way (compare with formula (3) in the third paper)

$$f(y_{S_1}|y_{S_2}, M_i) = \frac{\int L(\theta_i; y_{S_1}, M_i) L(\theta_i; y_{S_2}, M_i) \pi_i(\theta_i) d\theta_i}{\int L(\theta_i; y_{S_2}, M_i) \pi_i(\theta_i) d\theta_i}. \tag{6}$$

Depending on the form of the sets $S_1, S_2$, one can classify the available literature on BF in the following way (we do not list here all the possible cases)

(i) $S_1 = J_n,\ S_2 = \emptyset$ which yields the standard prior predictive or marginal density.

(ii) $S_1 = \{r\},\ S_2 = J_n - \{r\}$ which yields the cross-validation density $f(y_r|y_{(r)}, M_i)$ (compare with CPO model from the previous section).

(iii) (generalization of case(2)) $S_1$ is small subset of $J_n$; $S_2 = J_n \backslash S_1$.

(iv) $S_1 = J_n$; $S_2 = J_n$ which yields Aitkin's (1991) posterior predictive density.

Notice, that (i) yields the Bayes factor given by formula (5), (ii) and (iii) yield a so-called pseudo-Bayes factor (PSBF) (Geisser and Eddy (1979)), and finally (iv) yields a posterior Bayes factor (POBF) (see Aitkin (1991)).

## 7.3 A short note on asymptotics for PSBF

PSBF is the closest class of BFs to the one we would need to consider in performing model comparison. We stopped short by calculating predictive intensities using the MCMC algorithmic method. These posterior predictive intensities could naturally be used for calculation of appropriate BF.

Suppose that $g(\theta_i)$ is taken as an importance sampling density for $L(\theta_i; y_{S_2}, M_i)\pi_i(\theta_i)$. If $\theta^*_{ij},\ j = 1, ..., B_i$ is a sample from $g$ and we define $w_{ij} = L(\theta^*_{ij}; Y_{S_2}, M_i)\pi_i(\theta^*_{ij})/g(\theta^*_{ij})$. Then a Monte Carlo integration for density (6) yields

$$\hat{f}(y_{S_1}|y_{S_2}, M_i) = \Sigma_j L(\theta^*_{ij}; y_{S_1}, M_i) w_{ij} / \Sigma_j w_{ij}. \tag{7}$$

Let us follow the pattern used in the third paper and consider sample $\theta^*_{ij},\ j = 1, ..., B_i$ to be conveniently taken from the posterior $\pi_i(\theta_i|y)$. Then formula (7) gets the form

$$\hat{f}(y_{S_1}|y_{S_2}, M_i) = \{\Sigma_j \frac{1}{L(\theta^*_{ij}; y_{S_2^c}, M_i}\}^{-1} \Sigma_j \frac{L(\theta^*_{ij}; y_{S_1}, M_i)}{L(\theta^*_{ij}; y_{S_2^c}, M_i)}, \tag{8}$$

where $S_2^c = J_n - S_2$. It is a simple routine to calculate and the simulation is consistent. However, its precision depends on the stability of the weights $w_{ij} = L(\theta^*_{ij}; y_{S_2^c}, M_i)^{-1}$, i.e. on how good the importance sampling density $\pi_i(\theta_i|y)$ is for $\pi_i(\theta_i|y_{S_2})$. In the case when $S_2$ is small in comparison to $J_n$, we would expect $\pi_i(\theta_i|y)$ to be a good importance sampling density for each $\pi(\theta_i|y_{S_2})$. In the context of cross-validation ($S_2 = \{r\}$), equation (8) becomes a harmonic mean $\hat{f}(y_r|y_{(r)}, M_i) = B_i\{\Sigma_j \frac{1}{f_r(y_r|\theta^*_{ij}, M_i)}\}^{-1}$ from which PSBF can be calculated straightforwardly.

# 8 General Conclusions

My personal view is that statistical modeling provides much more vague or general answers than are usually asked for or expected by non-statisticians. The results of statistical analysis are reported in terms of estimates of parameters of interest and their confidence/credible intervals: the latter specifies the degree of uncertainty of the former. Parameters can be on population or on individual level.

Interpretation of results and their implications are often different on the individual and population levels. One can report clear quantitative findings on the population level, and at the same time only qualitative differences for selected individuals. For example, we report in the second paper that, in average, children attending nursery day care have several times higher probability of getting AOM than children at home day care. This finding does not imply that when we consider two kids, one who is at home day care and second attends nursery day care during their first two years of life, that the first one will get more AOM infections. This is exactly the kind of answers parents would be the most interested in. The answers they can get are only in terms of chances, not deterministic certainties.

Statistical inference has two potential sources of vagueness, i.e. accuracy of the entertained model, and approximation usually applied in order to get results. The applied models of this thesis describe the etiology of AOM, and inherit the above mentioned general sources of error of approximation. On the other hand, they are dynamic in time, non/semiparametric, and tested for goodness-of-fit. These features provide a full range of answers which statistical modeller could give.

In examining two models, it is clear that their predictive distributions will be comparable whereas their posteriors will not. They play rather complementary roles. Posterior is used for "estimation of parameters conditional on the adequacy of the model" whereas the predictive distribution is used for "criticism of the entertained model in light of the current data".

# References


Aalen, O.O. (1975), "Statistical Inference for a Family of Counting Processes," Ph.D. dissertation, University of California, Berkley.

Aitkin, M. (1991), "Posterior Bayes Factors (with discussion)" *J. of R.S.S., Soc.* B, **53**, 111-142.

Andersen, P.G., Borgan, Ø., Gill R.D. and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

Bayes, T. (1763), "An Essay Towards Solving a Problem in the Doctrine of Chances," *Biometrika* **45** (1958), 293-315.

Bernardo, J.M. and Smith, A.F.M. (1994), *Bayesian Theory*, Chichester: Wiley.

Box, G.E.P. (1980), "Sampling and Bayes Inference in Scientific Modelling and Robustness," *J. of R.S.S., Ser.* A, **143**, 383-430.

Cox, D.R. (1972a), "Regression Models and Life Tables," *J. of R.S.S., Ser.* B, **340**, 187-220.

Cox, D.R. (1972b), "The Statistical Analysis of Dependencies in Point Processes," *Stoch. Point Processes*, ed. P.A.Lewis, New York: Wiley, pp. 55-66.

de Finetti, B. (1974, 1975), *Theory of Probability* **1,2**, Chichester: Wiley.

Fleming, T.R. and Harrington, D.P. (1991), *Counting Processes and Survival Analysis*, Chichester: Wiley.

Geisser, S. and Eddy, W. (1979), "A Predictive Approach to Model Selection," *JASA*, **74**, 153-160.

Geyer, C.J. and Møller, J. (1994), "Simulation and Likelihood Inference for Spatial Point Processes," *Scand. J. Statist.* **21**, 359-373.

Ghosh, S.K., and Gelfand, A.E. (1995), "Model Choice: a Minimum Posterior Predictive Loss Approach," Technical report 95-06, Univ. of Connecticut, Dept. of Statistics.

Gilks, W.R., Richardson, S., Spiegelhalter, D. (1996) *et al.*, *Markov chain Monte Carlo*, Chapman&Hall.

Kass, R.E., and Raftery, A.E. (1994), "Bayes Factors," *JASA*, **90**, 773-795.

Lindley, D.V. (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint*, Cambridge: University Press.

Oakes, D. (1992), "Frailty Models for Multiple Event-Times," *Survival Analysis*, 371-379.

Oja, H., Alho, O.P. and Laara, E. (1996) "Model-based estimation of the excess fraction (attribute fraction): Day care and acute middle ear infection," *Statistics in Medicine*, **15**, 1519-1534.

Sinha, D. (1993), "Semiparametric Bayesian Analysis of Multiple Time Data," *JASA*, **88**, 979-983.

Sinha, D. and Dey, D.K. (1997), "Semiparametric Bayesian Analysis of Survival Data," *JASA*, **92**, 1195-1212.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *Ann. Statist.*, **22**, 1701-1786.

Tanner and Wang (1987), "The Calculation of Posterior Distributions by Data Augmentation (with discussion)," *JASA*, **82**, 528-550.

Vaupel, J.W., Manton, K.G., and Stallard, E. (1979), "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality," *Demography*, **16**, 430-454.

# Summaries of the original papers

**I A Note on Histogram Approximation in Bayesian Density Estimation** (Bayesian Statistics 5, eds. J.M. Bernardo, J.O.Berger, A.P. Dawid and A.F.M. Smith, Oxford University Press: 487-490, 1996). In Bayesian density estimation, it is in practice necessary to restrict the space of density functions in some way, in order to arrive at an effectively finite parametrization. Here we consider piecewise constant functions as an approximating family. We show that if such functions are used to support an approximation of the "true" prior, then, under a set of natural conditions, the corresponding approximation will hold for the posterior.

**II Acute Middle Ear Infection in Small Children: A Bayesian Analysis Using Multiple Time Scales** (*Lifetime Data Analysis (LIDA)*, **4**, 121-137, 1998). The study is based on a sample of 965 children living in the Oulu region (Finland), who were monitored for acute middle ear infections from birth to the age of two years. We introduce a nonparametrically defined intensity model for ear infections, which involves both fixed and time dependent covariates, such as calendar time, current age, length of breast-feeding time until present, or current type of day care. Unmeasured heterogeneity, which manifests itself in frequent infections in some children and rare in others and which cannot be explained in terms of the known covariates, is modelled by using individual frailty parameters. A Bayesian approach is proposed to solve the inferential problem. The numerical work is carried out by Monte Carlo integration (Metropolis-Hastings algorithm).

**III Predictive Inference, Causal Reasoning, and Model Assessment in Nonparametric Bayesian Analysis: A Case Study**, *Lifetime Data Analysis (LIDA)*, **6**, 187-205, 2000. This paper continues our earlier analysis of a data set on acute ear infections in small children, presented in Andreev and Arjas (1998). The main goal here is to provide a method, based on the use of predictive distributions, for assessing the possible causal influence which the type of day care will have on the incidence of ear infections. A closely related technique is used for the assessment of the nonparametric Bayesian intensity model applied in the paper. Two graphical methods, supported by formal tests, are suggested for this purpose.

**IV Joint Modelling of Recurrent Infections and Immune Response by Bayesian Data Augmentation**, submitted for publication in *JASA*, 2000. A joint dynamic model for the inter-dependence between infections, immune response and risk of disease is presented. We consider the recurrent subclinical infections as realisations from a renewal process and the antibody dynamics as a diffusion with decreasing drift modified by the effect of random infections. The augmented submodels are estimated simultaneously in one large Markov chain Monte Carlo algorithm. As an example, we consider the risk of recurrent ear infections (acute otitis media, AOM) caused by Streptococcus pneumoniae (Pnc) when having only partially observed information on mucosal colonisation and immune response. In particular, the protective role of antibodies induced by pneumococcal surface adhesin A (psaA) is studied. We found that, on average, every sixth pneumococcal carriage ends up with ear infection and in two thirds of such episodes, a preceding viral infection is present. Natural antibodies to PsaA are produced early as a reaction to infections but their protective effect is age-dependent, and even then only marginal.